## Title

Characterising communities impacted by the 2015 Indiana HIV outbreak: A big data analysis of social media messages associated with HIV and substance abuse

Permalink

https://escholarship.org/uc/item/5hg848zk

Journal

Drug and Alcohol Review, 39(7)

ISSN

0959-5236

Authors

Cuomo, Raphael E
Cai, Mingxiang
Shah, Neal
et al.

Publication Date

2020-11-01

DOI

10.1111/dar.13091

Peer reviewed

Published in final edited form as:
*Drug Alcohol Rev.* 2020 November ; 39(7): 908–913. doi:10.1111/dar.13091.

# Characterising communities impacted by the 2015 Indiana HIV outbreak: A big data analysis of social media messages associated with HIV and substance abuse

**RAPHAEL E. CUOMO**[1,2,3], **MINGXIANG CAI**[1,2,4], **NEAL SHAH**[1,2], **JIAWEI LI**[1,2], **WEN-HAO CHEN**[1,2,4], **NICK OBRADOVICH**[5], **TIM K. MACKEY**[1,2,3,6]

[1]Global Health Policy Institute, San Diego, USA

[2]Department of Healthcare Research and Policy, University of California, San Diego, USA

[3]Department of Anesthesiology, University of California, San Diego, USA

[4]Department of Computer Science and Engineering, University of California, San Diego, USA

[5]Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany

[6]Division of Infectious Disease and Global Public Health, University of California, San Diego, USA

## Abstract

**Introduction and Aims.—***Infoveillance approaches (i.e. surveillance methods using online content) that leverage big data can provide new insights about infectious disease outbreaks and substance use disorder topics. We assessed social media messages about HIV, opioid use and injection drug use in order to understand how unstructured data can prepare public health practitioners for response to future outbreaks.*

**Design and Methods.—***We conducted an retrospective analysis of Twitter messages during the 2015 HIV Indiana outbreak using machine learning, statistical and geospatial analysis to examine the transition between opioid prescription drug abuse to heroin injection use and finally HIV transmission risk, and to test possible associations with disease burden and demographic variables in Indiana and Marion County. Tweets from October 2014 to June 2015 were compared to disease burden at the county level for Indiana, and classification of census blocks by presence of relevant*

Correspondence to: Dr Timothy K. Mackey, Global Health Policy Institute, 8950 Villa La Jolla Drive, Suite A124, San Diego, CA 92037, USA. Tel: (951) 491 4161; tmackey@ucsd.edu.
Raphael E. Cuomo MPH, PhD, Investigator Global Health Policy Institute, Research Scientist, Department of Healthcare Research and Policy, and Assistant Professor, Department of Anesthesiology, Mingxiang Cai BA, Data Scientist, Global Health Policy Institute, Research Associate, Department of Healthcare Research and Policy, and Graduate Research Associate, Department of Computer Science and Engineering, Neal Shah BS, Research Associate, Jiawei Li MS, Data Scientist, Wen-Hao Chen BS, Research Associate, Global Health Policy Institute, Research Associate, Department of Healthcare Research and Policy and Graduate Research Associate, Department of Computer Science and Engineering, Nick Obradovich PhD, Senior Research Scientist and Principal Investigator, Tim K. Mackey MAS, PhD, Director, Global Health Policy Institute, Director, Department of Healthcare Research and Policy, Associate Professor, Department of Anesthesiology and Associate Professor, Division of Infectious Disease and Global Public Health.

Conflict of Interest
The authors have no conflicts of interest.

Supporting Information
Additional Supporting Information may be found in the online version of this article at the publisher’s website:

*messages was done at the census block level for Marion County. Marion County was used as it exhibited the highest total count of Tweets.*

**Results.—***257 messages about substance abuse and HIV were significantly related to HIV rates (*$P < 0.001$*) and opioid-related hospitalisations (*$P = 0.037$*). Using 157 characteristics from the American Community Survey, a linear classifier was computed with an appreciable correlation (*$r = 0.49$*) to risk-related social media messages from Marion County.*

**Discussion and Conclusions.—***Communities appear to communicate online in response to disease burden. Classification produced an accurate equation to model census block risk based on census data, allowing for high-dimensional estimation of risk for blocks with sparse populations.*

**Keywords**

human immunodeficiency virus; acquired immunodeficiency syndrome; social media; geospatial analysis; ecological epidemiology

## Introduction

Starting November 2014, the state of Indiana experienced an alarming outbreak of HIV infections among people who inject drugs in rural Scott County [1]. The first case attributed to this outbreak was recorded on 18 November 2014, rising to 17 new cases by 23 January 2015 [1]. By March, the HIV outbreak was declared a public health emergency by the State of Indiana and was attributed to 215 new HIV cases [1]. Subsequent studies have attributed the outbreak's origin to injection of the extended release opioid analgesic oxymorphone [2], especially among women exchanging sex for drugs [3]. Rural areas' tendency to lack syringe service programs, which provide sterile injection equipment, has also been implicated [4,5]. These factors imply the potential benefits of pre-emptive population-level surveillance for drug-related behaviours, especially those with high potential for additional negative outcomes like spread of infectious diseases.

Surveillance that leverages social media data can provide public health practitioners with information about social features unique to communities with risk-predisposing characteristics, thereby potentially elucidating more efficacious strategies to prevent morbidity and mortality from these outbreaks [6,7]. Indeed, prior research has indicated that 'infoveillance' approaches leveraging social media data can be beneficial for interventions seeking to target vulnerable populations as part of public health response to infectious disease outbreaks [8].

Linear classification can be a valuable tool to increase the effectiveness of public health programs. Strained funding for disease prevention programs must be allocated in a manner that objectively minimises morbidity and mortality. Classification identifies these at-risk communities using numerous characteristics from areas with relevant posts detected in this study. Therefore, by quickly identifying likely impacted communities with this imputed digital 'signal' of disease risk and elevated conversation, public health practitioners can prioritise preventive and health promotion communication efforts for areas using a more nuanced understanding of specific risks and concerns associated with these communities. This prioritisation may facilitate more effective public health programs, ultimately improving the use of funds to decrease morbidity and mortality lost from crises like the Indiana HIV epidemic of 2015. More broadly, an approach that leverages classification methods with social media and demographic data may be useful for public health practitioners seeking to discover communities which were previously unknown to be at high risk for appreciable morbidity/mortality from disease outbreaks.

In this study, we conduct a retrospective analysis of geocoded social media conversations on Twitter to provide insight into community discussion on topics associated with drug use and HIV during the 2015 Indiana HIV outbreak. To this effect, we conducted two layers of ecological analyses: (i) an assessment of the relationship between social media engagement and disease burden from HIV and opioid use; and (ii) a modelling exercise seeking to accurately classify areas that engage with these topics on social media in and near where an HIV outbreak occurred.

## Methods

The computer programming language Python, its associated packages and cloud Amazon Web Services were used to collect data for over 10 million messages from the Twitter API geolocated in the state of Indiana containing keywords related to opioids, intravenous drug use (IDU) and HIV. The data analysed were limited to a 7-month time frame (October 2014-June 2015), thereby capturing Tweets shortly before and during the HIV outbreak in Indiana. An unsupervised machine learning algorithm called the biterm topic model was used to identify and remove messages not related to opioids, IDU- or HIV-related user-generated messages. A final corpus of 1350 messages were manually reviewed by human coders ($\kappa$ = 0.94), resulting in 257 Tweets denoting behaviours, including drug usage, with relevance to the outbreak (see Appendix S1). Tweets not deemed relevant included news related messages about the outbreak and messages that did not originate from human users (such as commercial and bot traffic). Detailed analysis of all relevant messages detected in this analysis is being prepared in a separate study [9].

Messages having latitude/longitude coordinates were regressed with 2015 non-fatal opioid-related hospitalisations from data available from the Indiana State Department of Health and with new 2015 HIV cases from the US Centers for Disease Control and Prevention. These indicators of disease burden were taken from 2015 as they are expected to have been most impacted by the HIV outbreak, which peaked in early 2015. Analysis was done at the county-level with population-normalisation using estimates from the American Community Survey, and zeroes were imputed for values redacted due to low sample sizes. These analyses were done to elucidate if the geospatial distribution of Tweets from the time frame of the HIV outbreak was consistent with disease burden metrics influenced by this outbreak, while normalised for population.

To create an accurate classifier of social media engagement, 3865 variables at the census block level were obtained from the American Community Survey’s 5-year (2011–2015) using the American FactFinder website (factfinder.census.gov), which allows for users to search data collected by the US Census Bureau. These variables were the extent of available information at the census block level; they are routinely collected information available to public health agencies that can be used to better understand populations of small areas, and this modelling analysis proceeds in a manner concordant with information which would have been available to public health agencies during the HIV outbreak. Furthermore, the 2011–2015 5-year estimate was specifically chosen as later estimates would not have been available to public health agencies during the time frame of the HIV outbreak. Though the epicentre of the outbreak was Scott County, Marion County (population 950 000) was selected for this analysis because it was the only county exhibiting a sufficient number of Tweets, and only four Tweets were collected from Scott County (population 24 000). We removed duplicate measures, measures with missing data, variables with zeroes and the lower half of Pearson's correlation coefficients with message frequency. Loadings from linear discriminant analysis—unitless measures of correlation between independent predictors and the dependent variable, adjusting for the influence of other predictors—were used to combine 157 characteristics to compute a model that uses routinely collected demographic data to accurately predict the occurrence of relevant Tweets at the census block level in Marion County. Values were visualised as a choropleth map using ArcGIS. The influence of spatial autocorrelation on Tweet occurrence by census tract was assessed by computing the Morans I statistic.

## Results

Reports of HIV and IDU were highly prevalent among the 257 relevant Tweets assessed in this study. The following are examples of Tweets which suggest opioid abuse: 'Bout to roll up take these percs and sip this drink, icant deal', 'they gave me hydros earlier l o l', 'High off these percs', 'I'm gnoig home and ovredsoing on heroin and cOCaine', 'hope that heroin bitch was worth it', 'My mom just asked me for happy pills and alcohol?', 'I GOT REALLY DRUNK ALL DAY YESTERDAY, I DID COCAINE AND HEROIN LOL'.

Half ($n = 46$) of Indiana's counties were included in 257 relevant Tweets. There was a statistically significant relationship between per capita messages and per capita opioid-related hospitalisations ($\beta = 1.073$, $P = 0.037$, $R^2 = 0.047$) and with per capita new HIV cases ($\beta = 10.105$, $P < 0.001$, $R^2 = 0.24$). Figure 1 illustrates data used for county-level analyses.

Eighty-six Tweets were detected for 62 of Marion County's 632 census blocks, the most populous area of the state. Spatial autocorrelation did not appear to exert a statistically significant influence on Tweet presence ($I = 0.041$; $P = 0.056$). Concordantly, Figure 2 suggests high heterogeneity of relevant messages. Though the 157 characteristics chosen for combination into the classifier appear to be closely related to the total population. The classifier exhibited an $r = 0.49$ with relevant Tweets per capita, whereas total population exhibited an $r = 0.13$.

## Discussion

Twitter interaction pre-and post-outbreak on HIV, opioid and IDU-related topics was significantly associated to HIV and opioid burden in separate models utilising population-adjusted covariates. These findings suggest that communities engage in online communication in reaction to their community-specific disease burdens, consistent with findings from studies assessing the relationship between Tweets and other infectious diseases, including influenza [10], pertussis [11] and conjunctivitis [12], as well as abuse of alcohol [13] and amphetamines [14].

This classification exercise used the presence of posts and their spatially linked census data to uncover additional areas at potentially high risk for HIV/opioid-related communication and behaviour. These areas include a suburban community southwest of Sunshine Gardens, areas around Eagle Creek Park, and the towns of Wynnedale and Spring Hill north of Indianapolis. As individuals may be apprehensive about public messaging on HIV/opioid issues, classification allows the use of characteristics for areas with available messages to infer risk for areas without available messages.

Our analyses indicate that HIV, opioid and IDU themes may be attributable to youth and adolescent substance abuse discussion, as numerous messages were detected near schools. Additional studies are needed to confirm whether significant correlations uncovered in this study are indicative of a causal relationship. In particular, further studies should assess whether relaxed norms for opioid prescription drug and polydrug abuse are related to discussion of heroin and IDU, as indicated by the previously noted example messages.

Communities afflicted by the opioid crisis face a potentially dangerous transition to heroin use, thereby increasing the risk of communicable disease transmission, including from HCV/HIV. For Scott County, Indiana, measures designed to prevent HIV transmission, such as syringe exchange programs and HIV testing were unavailable and worsened the progression of the outbreak [15]. Importantly, earlier detection of risk factors preceding the outbreak and a swifter response from Scott County and the State of Indiana could have been implemented based on additional data [1]. Therefore, there is utility in overlaying readily available census data with social media messages, as was done in this study.

Limitations of this study include muted generalisability to states with markedly different demographic characteristics, especially with respect to age distributions and macroeconomic characteristics. These analyses, particularly at the county level, suffer from ecological fallacy, so care should be taken in extrapolating to individual behaviour. Further research is needed to assess validity of these observations and real-world predictive value in detecting and responding to past and future HIV outbreaks associated with injection drug use. The Twitter platform relays posts tied to users' account handles and profile pictures, which may result in apprehension to post about relatively sensitive activities such as drug use and HIV status. Nevertheless, this form of infoveillance does permit geolocation of posts, thereby allowing for data about community characteristics to be compared between areas with posts and those without posts. However, Tweets implying drug use may be tied to geospatial coordinates consistent with the location of drug use, even though demographic data are tied to location of residence. Also, individuals tweeting about observed drug use may not have used drugs themselves.

## Conclusion

In conclusion, results from this study suggest that statistical methods which leverage publicly available data at resolute geospatial levels can efficaciously be deployed as part of future HIV outbreak surveillance and possible prevention efforts. Furthermore, these strategies may lead to better understanding of at risk populations and communities for transition between opioid abuse and HIV-risk related IDU behaviour.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgement

The authors received an AIDS Research Development Grant from the UC San Diego Center for AIDS Research (NIAD 5 P30 A1036214) in support of this study.

## References

[1]. Gonsalves GS, Crawford FW. Dynamics of the HIV outbreak and response in Scott County, IN, USA, 2011–15: a modelling study. Lancet HIV 2018;5:e569–77. [PubMed: 30220531]

[2]. Conrad C, Bradley HM, Broz D et al. Community outbreak of HIV infection linked to injection drug use of oxymorphone—Indiana, 2015. MMWR Morb Mortal Wkly Rep 2015;64:443–4. [PubMed: 25928470]

[3]. Peters PJ, Pontones P, Hoover KW et al. HIV infection linked to injection use of oxymorphone in Indiana, 2014–2015. N Engl J Med 2016; 375:229–39. [PubMed: 27468059]

[4]. Campbell EM, Jia H, Shankar A et al. Detailed transmission network analysis of a large opiate-driven outbreak of HIV infection in the United States. J Infect Dis 2017;216:1053–62. [PubMed: 29029156]

[5]. Jarlais DCD, Nugent A, Solberg A, Feelemyer J, Mermin J, Holtzman D. Syringe service programs for persons who inject drugs in urban, suburban, and rural areas—United States, 2013. MMWR Morb Mortal Wkly Rep 2015;64:1337–41. [PubMed: 26655918]

[6]. Young SD. A ‘big data’ approach to HIV epidemiology and prevention. Prev Med 2015;70:17–8. [PubMed: 25449693]

[7]. Spiller MW, Broz D, Wejnert C et al. HIV infection and HIV-associated behaviors among persons who inject drugs—20 cities, United States, 2012. MMWR Morb Mortal Wkly Rep 2015;64: 270–5. [PubMed: 25789742]

[8]. Charles-Smith LE, Reynolds TL, Cameron MA et al. Using social media for actionable disease surveillance and outbreak management: a systematic literature review. PLoS One 2015;10:e0139701. [PubMed: 26437454]

[9]. Cai M, Shah N, Li J et al. Identification and characterization of Tweets related to the 2015 Indiana HIV outbreak: a retrospective infoveillance study, 2020. (Under review).

[10]. Aslam AA, Tsou M-H, Spitzberg BH et al. The reliability of Tweets as a supplementary method of seasonal influenza surveillance. J Med Internet Res 2014;16:e250. [PubMed: 25406040]

[11]. Nagel AC, Tsou M-H, Spitzberg BH et al. The complex relationship of real space events and messages in cyberspace: case study of influenza and pertussis using Tweets. J Med Internet Res 2013;15:e237. [PubMed: 24158773]

[12]. Qiu R, Hadzikadic M, Yu S, Yao L. Estimating disease burden using Internet data. Health Informatics J 2019;25:1863–77. [PubMed: 30488754]

[13]. Hossain N, Hu T, Feizi R, White AM, Luo J, Kautz H. Inferring fine-grained details on user activities and home location from social media: detecting drinking-while-tweeting patterns in communities. arXiv, arXiv: 1603.03181, 2016.

[14]. Hanson CL, Burton SH, Giraud-Carrier C, West JH, Barnes MD, Hansen B. Tweaking and tweeting: exploring Twitter for nonmedical use of a psychostimulant drug (Adderall) among college students. J Med Internet Res 2013;15:e62. [PubMed: 23594933]

[15]. Rich JD, Adashi EY. Ideological anachronism involving needle and syringe exchange programs: lessons from the Indiana HIV outbreak. JAMA 2015;314:23–4. [PubMed: 26000661]

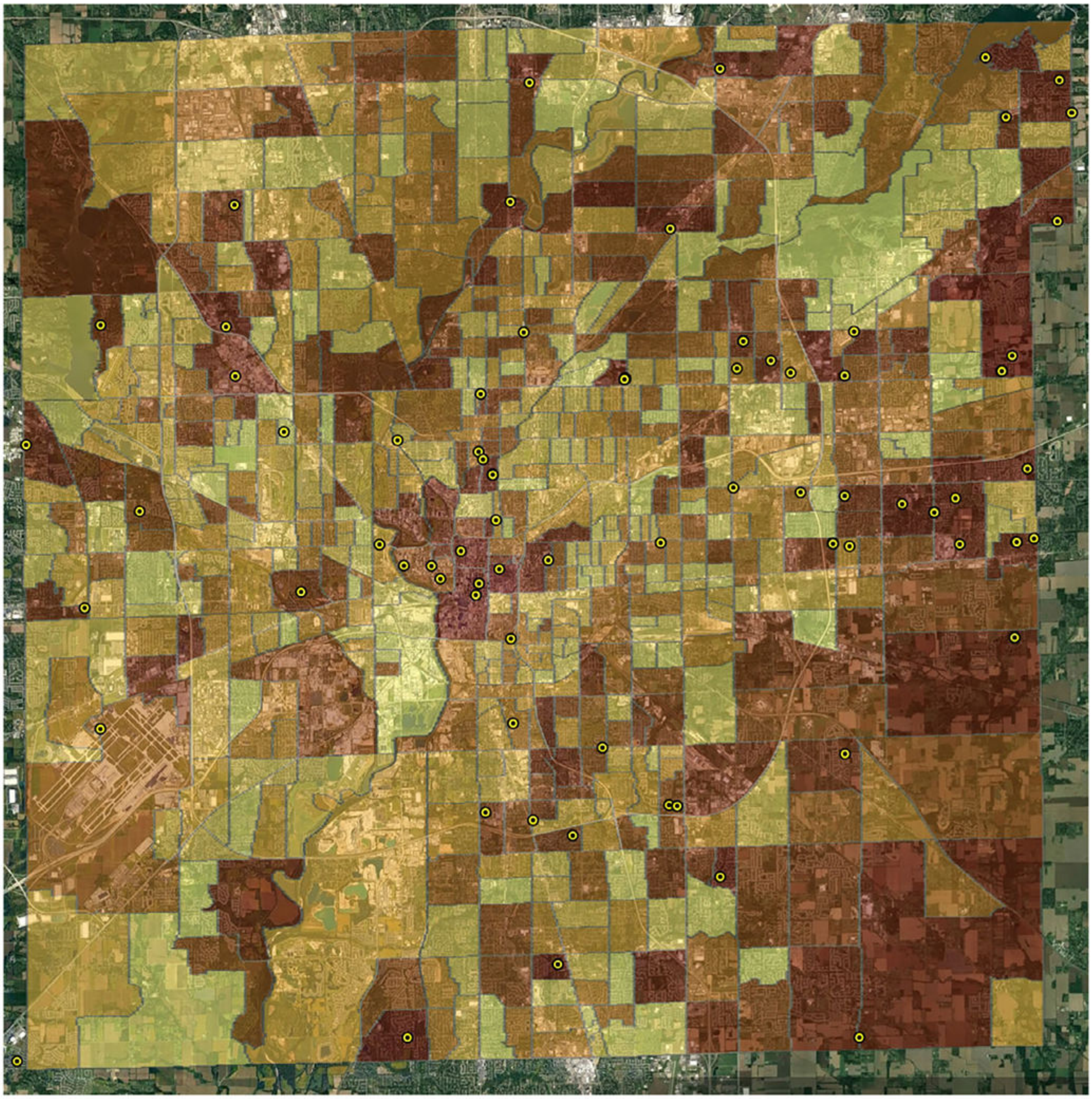

**Figure 1.**

Observed Tweets, 2015 non-fatal opioid-related hospitalisations and new 2015 HIV cases per 100 000 population, for all Indiana counties, denoted in a choropleth gradient. Values of zero were imputed for data points redacted due to low sample sizes, and these values are illustrated in grey.

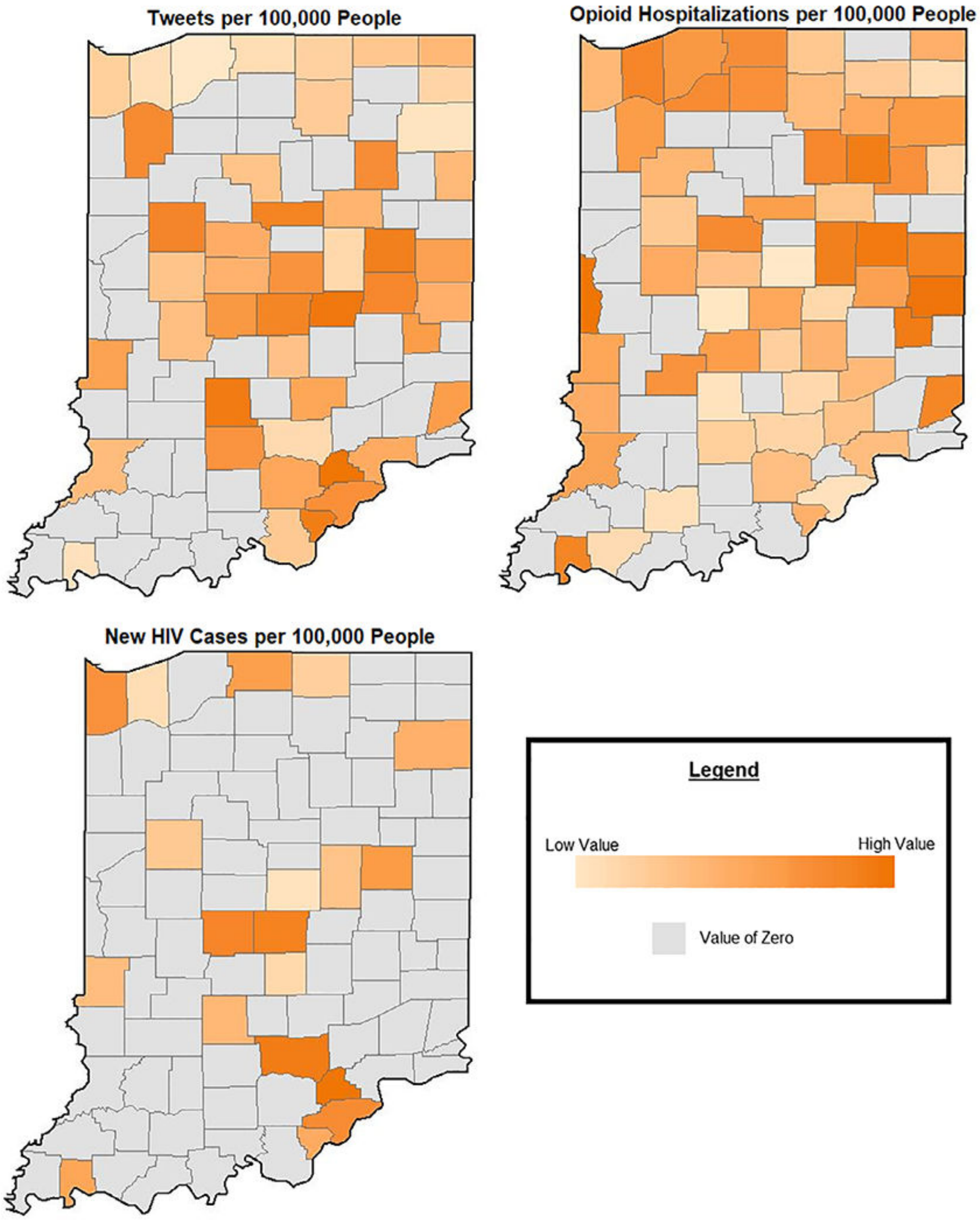
Tweets per 100,000 People
Opioid Hospitalizations per 100,000 People
New HIV Cases per 100,000 People
Legend
Low Value
High Value
Value of Zero

**Figure 2.**

Classification of 632 census blocks in Marion County, Indiana, for social media engagement with HIV/opioid/intravenous drug use themes during (and immediately pre/post) the 2015 HIV outbreak, computed using 157 demographic characteristics from the American Community Survey, with redder shades indicating higher propensity for social media engagement and points indicating observed Twitter posts on these topics.